ITC at DIMEMEX: When hate goes Viral: Detection of Hate Speech in Mexican Memes Using Transformers

Ramón Zatarain Cabada[1†], María Lucía Barrón Estrada[1†], Ramón Alberto Camacho Sapien[1†], Víctor Manuel Bátiz Beltrán[1†*], Néstor Leyva López[1†], Manuel Alberto Sotelo Rivas[1†]

[1] *Tecnológico Nacional de México Campus Culiacán, Culiacán, Sinaloa, México*

Abstract

This article presents the work done in the task of detecting abusive content in memes through the use of images and text, in the DIMEMEX contest as part of IberLEF 2024. Like any violent event, memes with hate speech that circulate through the network generate a negative impact on society, affecting not only the people directly involved in their creation or spreading, but also vulnerable groups and the health of the social fabric in general. Precisely, our participation focused on making use of the dataset provided by the organizers to perform the task of detecting hate speech (or "toxicity") in memes using visual-textual information. To solve the contest task an approach focused on the use of OCR and Transformers was used. Our proposal was based on BETO model and obtained, for subtask 1, an f1-score value of 0.48, ranking fourth place in the final phase. We conclude that this task is very complicated, but we consider that our results and others are promising.

Keywords

Hate Speech, Sentiment Analysis, NLP, LLM, Deep Learning, Transformers, Machine Learning

1. Introduction

In everyday language, hate (or toxic) speech refers to offensive discourse of a discriminatory or pejorative nature, targeting a group or individual because of inherent characteristics or "identity factors" (such as race, religion, or gender) and which may threaten social peace. This type of discourse can be transmitted through any form of expression, including images, caricatures, cartoons, objects, gestures, symbols and even memes [1]. With respect to the latter, as defined by Oxford, "it is an image, video, piece of text, etc. typically humorous in nature, which is copied and spread rapidly by internet users,

IberLEF 2024, September 2024, Valladolid, Spain

* Corresponding author.

† These authors contributed equally.

ramon.zc@culiacan.tecnm.mx (R. Zatarain); lucia.be@culiacan.tecnm.mx (M. L. Barrón); ramon.cs@culiacan.tecnm.mx (R. A. Camacho); victor.bb@culiacan.tecnm.mx (V. M. Bátiz); nestor.ll@culiacan.tecnm.mx (N. Leyva); manuel.sr@culiacan.tecnm.mx (M. A. Sotelo)

0000-0002-4524-3511 (R. Zatarain); 0000-0002-3856-9361 (M. L. Barrón); 0009-0003-9367-7730 (R. A. Camacho); 0000-0003-4356-9793 (V. M. Bátiz); 0000-0002-2767-5708 (N. Leyva); 0009-0008-5879-871X (M. A. Sotelo)

often with minor variations." That said, with the ease of creating humorous satire with a piece of multimedia, a large number of these are likely to be offensive in nature. Given this, it has become an important issue for researchers around the world to identify memes with hate speech. This is because it would be easier to act against the bullying and discrimination that vulnerable people often suffer, since these memes tend to spread harmful ideas and messages that incite violence against minority groups. It can also be a useful tool to educate society about the dangers of hate speech and how to recognize it and respond to it. This paper presents the work carried out in the participation in the DIMEMEX competition [2], as part of IberLEF 2024 [3], in the task of detecting toxicity or violence in memes in Spanish. For this, an analysis of the provided dataset was performed, identifying the distribution of the data, as well as the categorization of these. In addition, operations corresponding to data cleaning and processing were performed to eliminate non-relevant content in the data. Subsequently, several Natural Language Processing (NLP) and Optical Character Recognition (OCR) models were used, screened and evaluated, including models based on Transformers. As a product of this work, the results and conclusions obtained are shown.

2. Related Work

Recognizing toxicity in texts presented in memes is closely related to understanding the context of the image and the writing in it, so that it can be understood if the comments present toxicity. As reported by Xenos et al. [4], it raises the question of how context influences human judgment and improves performance in toxicity systems. Using a dataset obtained from Wikipedia, it was found that context can amplify or mitigate the perceived toxicity of messages. A similar case is presented by Rupapara et al. [5], as it considers that the identification of toxic comments is essential in social networks. For this purpose, they used a Regression Vector Vote Classifier (RVVC). The result of their proposed method suggests that their F1 score results outperforms other individual models when using features with a balanced dataset, reaching an accuracy of 0,97. A similar scenario is presented by the work of Wang et al. [6], which aims to create a toxicity detector in texts extracted from the Internet using machine learning methods (CNN, Naive Bayes model and LSTM). The objective is to build such models to provide a higher accuracy of the predecessors. After the experiments, it was obtained that the LSTM network achieves the highest accuracy, recognizing that there is an opportunity for improvement in the preprocessing part. Going deeper into the use of machine learning networks to detect toxic speech, the work of Malik et al. [7], considering that hate speech and offensive content increased exponentially with the COVID-19 pandemic, set out to create a dataset with comments of this nature. After preprocessing it through NLP and embeddings (making use of BERT), and using various deep learning networks, CNN gave the best result, as it can adapt to understand and identify the right patterns in word sequences.

It is important to emphasize that not all writings are in text format, but a large part of them is found as text within images, so it is vital to have the necessary tools to recognize text within images. Due to this pressing need, work such as that of Xue et al. [8] presents supervised pre-training methods for acquiring effective representations of text in images by jointly learning and aligning visual and textual information. Relying on a network with

an image encoder and a text encoder with character recognition that extract visual-textual features. Their experiments show that such a pre-trained model improves the F1 score by more than 2%, and by more than 4% when transferring its weights to other text detection and localization networks. A similar case is studied by Liao et al. [9], where a Differentiable Binarization (DB) module is proposed to perform segmentation in a network. Since a segmentation network can adaptively set binarization thresholds, which not only simplifies post-processing, but also improves text detection performance. Achieving competitive performance with different datasets at real-time speed.

This type of work is fundamental to find hate speech found in memes in different social networks. Previous work has been carried out on this topic, such as the one presented by Suryawanshi et al. [10]. Where it is deemed as necessary to combine the modality of text and image, in addition to the context, to identify whether memes are offensive or not. For this purpose, a classifier was developed to detect offensive content in a dataset designed with this premise. In addition, an early fusion technique was used to combine the image and text modality to compare with a text baseline and an image-only baseline. Resulting in improvements in precision, recall and F1 score. Similarly to Chen et al. [11], whose study aims to get closer to detecting hate messages in memes. To achieve this objective, a triplet was fed by stacking visual features, object labels, and text features of memes generated by a Visual Features in Vision-Language (VinVI) detection model and Optical Character Recognition (OCR) technologies. Demonstrating that data with anchor point addition can improve deep learning-based toxic meme detection performance by involving more substantial alignment between text caption and visual information.

The difference between the work presented in this paper and previous work is the focus on using OCR and the BETO Transformer model in combination with LLM models. In this way, an integrated work system is created between different models, aiming to cover most aspects of NLP from images.

3. Task Description

The competition explained in detail in [2, 12] was divided into two tasks:
Subtask 1 is about identification of presence of one of three classes: hate speech, inappropriate content, and neither.
Subtask 2 is about using a finer-grained classification to distinguish instances of hate speech into different categories, including classism, sexism, racism, and others.
For the evaluation of the subtask solution proposals, the competition organizers established that macro-average of precision, recall and f1-score will be the leading evaluation measures for both subtasks. The Codalab platform [12, 13] was used for the submission of proposals and their evaluation.

4. Methodology

4.1. Dataset Description

The organizers of the competition provided a training dataset with 2263 records, each record containing two fields, one of them to indicate the name of the meme image and the other one for the text related to the meme. In addition, two files were provided containing the labels corresponding to each of the two tasks. A link to download the images corresponding to the memes is also provided. A sample of the dataset is shown in Table 1.

Table 1
Head of the dataset

Image Filenames	Text
DS_IMG_2973	Quien se comió la ultima rebanada de pastel?! Yo: no lo sé preguntarle a otro Mi mamá: *le pregunta ; mi hermano quien lo hizo* El: acusa* Yo: No es sierto no seas mentiroso *me
DS_IMG_574	"No seas nena "Ser niña no significa ser cobarde 0 débil. No la uses como sinónimo aeiee @CONAVIM_MX CONAVIM CONAVIM WWW gob.mx/conavim
DS_IMG_263	PUEDE ZZEVARSE @MOR PURO VERDADÉRO POR TODA LA ETERNIDAD COMO NoS MUESTRA EL ADORABLE SMITHERS FEHNET DnnNci 0 PUEDE @@MIBIARLO POR ESTA CAJA DE BRANCA DE LITRO
DS_IMG_119	Peroyonoescriboen Wättpad FB: An INFP Mind Carajo eres INFP; escribes enWattpad
DS_IMG_624	Ysientes por mi ? No siento ni la alarma y voy a sentir algo por ti algo

4.2. Dataset Analysis

The analysis of the data set was an important starting point for the development of this work. The objective of this stage was to list aspects of the dataset that could negatively affect the training of the models.
To do this, we took random samples from the dataset where we analyzed the extracted text and compared it with the actual text in the image. The result of this process led us to the conclusion that several of the extracted texts corresponded partially with the text shown in the image (See Figure 1). The inconsistencies we found in the data set are listed below:

- The extracted text did not contain the words present in the image.
- The extracted text contained all the words present in the image, but not in the order presented.

- The extracted text included usernames, Facebook group names or links to web pages resulting from the text extraction.
- The extracted text contained special characters not present in the image.

These conclusions do not seek to generalize the data present in the dataset but rather tell us about particularities observed in the dataset, which could affect the training of the models and which, if solved, could lead to better results. The following sections describe the approaches used by the research team to address these issues.

Figure 1. Two images are shown in the figure. In the left image, the extracted text includes a username. In the image on the right, the extracted text is not in the original order.

4.3. Data Preprocessing

In order to resolve the inconsistencies listed in the previous section, several preprocessing techniques were applied. A detailed diagram involving all the processes applied to the data set to prepare it for model training is shown in **Figure 2**.

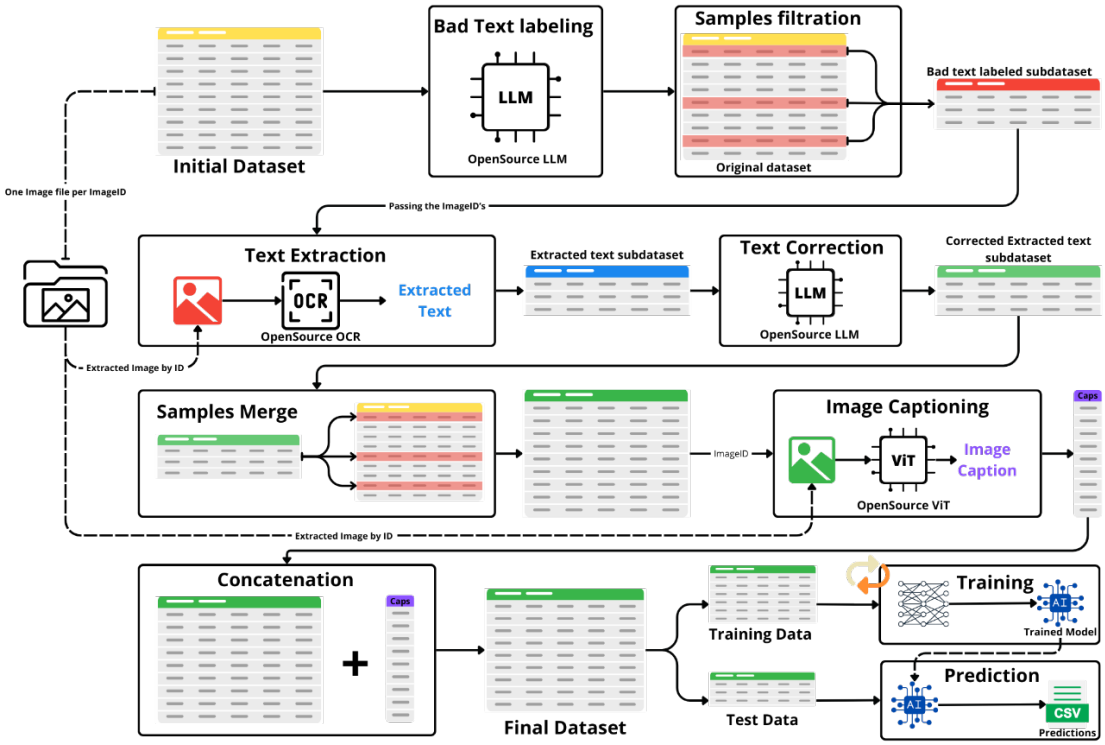

Figure 2. Process for preparing the data set to be used to train the classifier.

First, we started with the provided data set, which contained the text and the image. Then, using an open-source model, an analysis of the texts was made to determine if the text had coherence, if it did not present coherence, these data were separated into a small data set to apply another processing. Once the data detected as inaccurate was obtained, the text was extracted again from the image by applying an OCR, resulting in another small data set with the text and the image. A correction process was then applied to the texts using an open source LLM, and the corrected texts were integrated into the full dataset. Finally, each image was processed using the BLIP (Bootstrapping Language-Image Pre-training) ViT model to obtain its textual description, and the resulting texts were concatenated with the dataset to form the final dataset.

4.3.1. Labeling of incorrect texts

It was necessary to identify the texts in the dataset that were considered incorrect under the items listed in section 4.2. Due to the sample size of the dataset, it was decided to perform this procedure automatically. To achieve this, LlaMA [14] an open-source Large Language Model (LLM) was used. This model was assigned the task of identifying erroneous texts in the dataset by giving specific instructions through a “prompt” directly to the “System” user of the model. The model labeled each sample containing a text considered

“erroneous”. The samples labeled as erroneous were then extracted in a subset to be treated in the following steps.

4.3.2. Text Extraction using an OCR and Text correction using a LLM

Using the subset of inaccurate samples, we iterated over each image using the image ID obtained from the "Image_ID" column present in the dataset. A text extraction process was applied to each image using PaddleOCR [15] which, being OpenSource, allowed to replicate the experiment with no complications. At the end of this process, a subset of the wrong samples was extracted, now with a completely new extracted text.

Due to the complexity of the previous task, the extracted text from the image by OCR required corrections. These corrections were minimal and mostly dealt with words joined with other words. To perform these corrections automatically, the LLaMA model was used as well, but now with a different prompt. The assigned prompt specified that, from the extracted text of the image, identify the joined words and then separate them. Also, it was explicitly specified not to insert new words into the extracted text. In this process, the model received as input a text with inconsistencies (for example, joined words) and as output it generated a corrected text, as seen in **Figure 3**.

After this process concluded, a subset with clean and clear extracted texts was obtained from the image. This subset was integrated to the original data set, replacing the original samples with the new ones from the corrected subset.

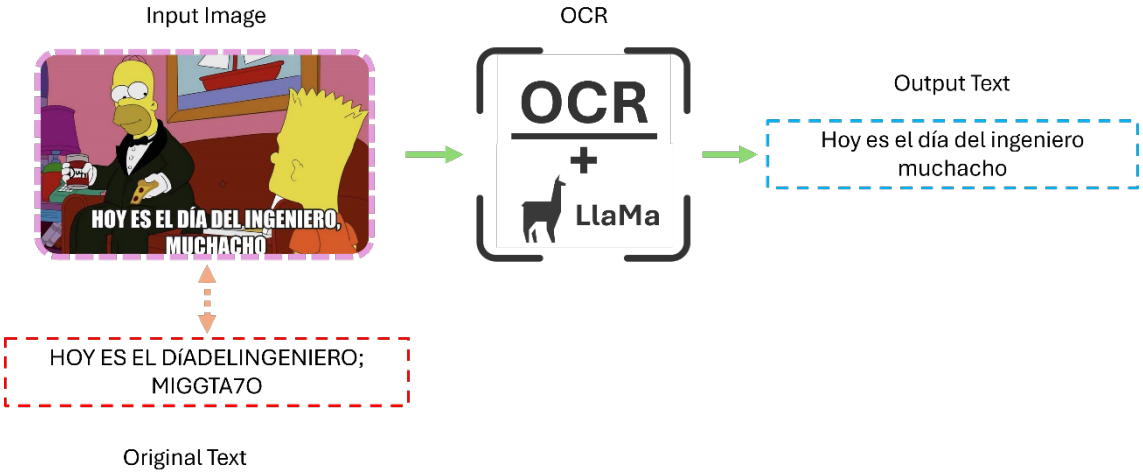

Figure 3. Corrected text extraction using OCR and LlaMa.

To further expand and enhance the dataset, a new feature was added, which was a text description of the image to each sample. The image description was done using a ViT-based model [16] configured specifically for this task where the model receives an image as input and provides text describing the input image as output. This description gives the flexibility to compare a model trained only with the texts extracted from the image. In addition, a

trained model was created with these texts in combination with the text resulting from the image description.

4.4. Model Selection and Training

Model selection and training are critical steps in machine learning, as it is used to find the optimal model that accurately represents the data and makes reliable predictions.

To perform this task, as a first alternative, a BERT-based model was used. As second, a data classification method was proposed based on the BETO [17] neural network, which is an adaptation of the BERT architecture designed for Natural Language Processing (NLP) tasks in Spanish. Subsequently, three independent models were trained within this same network (one for each target label). Taking these labels as a reference, their individual probabilities were combined with the model metrics. The performance of the model was evaluated using F1-score. In this way, the final labels were assigned to each data point. The label with the highest value was assigned a value of "1", while the remaining two labels were assigned a value of "0".

4.5. Impact of Data Enhancement and Captions from images Usage on Text Classification Model

In this section, we present a comparative analysis of text classification models trained on original versus enhanced datasets, as well as the impact of using the captions extracted from the images. The evaluation metrics analyzed include evaluation loss, accuracy, F1 score, and recall.

Below is the Table 2 summarizing the evaluation metrics for models trained on original versus improved datasets, and with versus without captions:

Table 2
Summary of Evaluation Metrics for Text Classification Models Trained on Original vs. Improved Data and With vs. Without Captions for subtask 1.

Label	Data	Using Captions	Accuracy	Recall	F1	Recall
Hate Speech	Improved	True	1.167	0.832	0.819	0.832
Inappropriate	Improved	True	1.372	0.779	0.759	0.779
Harmless	Improved	True	2.010	0.686	0.680	0.686
Hate Speech	Improved	False	1.175	0.850	0.835	0.850
Inappropriate	Improved	False	1.677	0.752	0.736	0.752
Harmless	Improved	False	0.633	0.659	0.657	0.659
Hate Speech	Original	False	0.940	0.834	0.806	0.834
Inappropriate	Original	False	0.557	0.801	0.739	0.801
Harmless	Original	False	1.810	0.647	0.640	0.647
Hate Speech	Original	True	1.205	0.795	0.793	0.795
Inappropriate	Original	True	1.056	0.784	0.759	0.784
Harmless	Original	True	2.384	0.664	0.659	0.664

4.5.1. Evaluation Metrics: Original vs. Improved Data

Models trained on improved data consistently exhibit higher accuracy across all labels (*hate speech*, *inappropriate*, *harmless*). This indicates that data enhancement has a significant positive effect on model performance, leading to more accurate predictions. The evaluation loss is significantly lower for models trained on improved data compared to those trained on original data. This reduction in loss suggests that the models with improved data are better at generalizing from the training set to unseen data. Both the F1 score, and recall are higher for models trained on improved data. These metrics indicate that enhanced data not only improves the precision of the models but also their ability to correctly identify positive cases (i.e., true positives).

4.5.2. Evaluation Metrics: Captioned vs. Non-Captioned

Models trained with captions added to the original text tend to show a slight improvement in accuracy and a lower evaluation loss compared to those without captions. This suggests that captions provide additional contextual information that helps the model make more accurate predictions. The inclusion of captions from the images in the text also improves the F1 score and recall, indicating a more robust performance in correctly classifying both positive and negative cases.

5. Results

For both phases of the competition (development and final), the data set provided was divided into 80% for training and 20% for validation. The results obtained are presented below.

The proposed solution was built using the dataset given for the development phase and published on the Codalab platform. Our model was applied on the unlabeled dataset provided for the final phase. The proposed solution was generated in a CSV formatted file and uploaded for evaluation to the Codalab platform. Initially, we used a model based on BERT, but later, we got better results with a model based on BETO. This approach received a score of 0.48 on the f1 measure, placing our proposal as fourth place in the final phase of the competition, as shown in Table 3 (our submissions were made under the team's name ITC).

Table 3
Subtask 1 Final phase results

#	User/Team	f1	Precision	Recall
1	CLTL	0.58 (1)	0.61 (2)	0.56 (1)
2	CUFE	0.56 (2)	0.63 (1)	0.53 (2)
3	aaman	0.49 (3)	0.49 (4)	0.49 (4)
4	**ITC**	**0.48 (4)**	**0.48 (5)**	**0.47 (5)**
5	fariha32	0.47 (5)	0.52 (3)	0.50 (3)

6	mashd3v	0.42 (6)	0.48 (6)	0.42 (6)
7	CyT_Team	0.36 (7)	0.36 (7)	0.36 (7)
8	hugojair	0.27 (8)	0.31 (8)	0.31 (8)

6. Conclusions

This paper presents the participation in the DIMEMEX contest as part of IberLEF 2024 in the classification of abuse content in memes in Mexican Spanish. It was decided to participate only in subtask 1. The best result was obtained using a pre-trained model based on BETO with which the team's proposal came in fourth place in the final phase under the macro-average f1-score metric.

The analysis reveals that enhanced data significantly improves model performance across all evaluation metrics by providing richer and more diverse information, leading to better generalization and reduced overfitting. Additionally, the inclusion of captions further boosts the F1 score and recall, enhancing the model's ability to accurately classify both positive and negative cases. This added contextual richness is particularly useful for tasks requiring nuanced understanding, such as detecting hate speech or inappropriate content.

In conclusion, the combination of data enhancement and the use of captions, leads to superior text classification models. These findings highlight the importance of high-quality, context-rich training data in developing robust and reliable machine learning models. Future work could explore further refinement of data enhancement techniques, the integration of additional contextual information, and the use of data augmentation techniques to continue improving model performance.

Acknowledgements

We want to express our gratitude to CONAHCYT and the Tecnológico Nacional de México campus Culiacán for supporting our team to participate in the DIMEMEX@IberLEF 2024 challenge for detecting abuse content in Mexican Spanish memes.

References

[1] United Nations. Understanding Hate Speech. 2023. Retrieved June 11, 2024, from https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech.

[2] H. Jarquín Vásquez, I. Tlelo-Coyotecatl, D. I. Hernández Farías, M. Casavantes, H. J. Escalante, L. Villaseñor-Pineda, M. Montes y Gómez. Overview of DIMEMEX at IberLEF 2024: Detection of Inappropriate Memes from Mexico. Procesamiento del Lenguaje Natural, September, 2024.

[3] L. Chiruzzo, S. M. Jiménez-Zafra, F. Rangel. In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS.org

[4] A. Xenos, J. Pavlopoulos, I. Androutsopoulos, L. Dixon, J. Sorensen, L. Laugier. Toxicity detection sensitive to conversational context. First Monday, 2022.
[5] V. Rupapara, F. Rustam, H. F. Shahzad, A. Mehmood, I. Ashraf, G. S. Choi. Impact of SMOTE on imbalanced text features for toxic comments classification using RVVC model, 2021. IEEE Access, 9, 78621-78634.
[6] K. Wang, J. Yang, H. Wu. A survey of toxic comment classification methods, 2021. https://doi.org/10.48550/arXiv.2112.06412
[7] P. Malik, A. Aggrawal, D. K. Vishwakarma. Toxic speech detection using traditional machine learning models and bert and fasttext embedding with deep neural networks. In 2021 5th International Conference on Computing Methodologies and Communication (ICCMC) (2021, April) (pp. 1254-1259). IEEE.
[8] C. Xue, W. Zhang, Y. Hao, S. Lu, P. H. Torr, S. Bai. Language matters: A weakly supervised vision-language pre-training approach for scene text detection and spotting. In European Conference on Computer Vision (2022, October) (pp. 284-302). Cham: Springer Nature Switzerland.
[9] M. Liao, Z. Wan, C. Yao, K. Chen, X. Bai. Real-time scene text detection with differentiable binarization. In Proceedings of the AAAI conference on artificial intelligence (2020, April) (Vol. 34, No. 07, pp. 11474-11481).
[10] S. Suryawanshi, B. R. Chakravarthi, M. Arcan, P. Buitelaar. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In Proceedings of the second workshop on trolling, aggression and cyberbullying (2020, May) (pp. 32-41).
[11] Y. Chen, F. Pan. Multimodal detection of hateful memes by applying a vision-language pre-training model. Plos one, 2022, 17(9), e0274300.
[12] DIMEMEX. Challenge Website. (2024). Retrieved June 11, 2024, from https://codalab.lisn.upsaclay.fr/competitions/18118.
[13] A. Pavao, I. Guyon, A. Letournel, D. Tran, X. Baró, H. J. Escalante, S. Escalera, T. Thomas, Z. Xu. CodaLab Competitions: An open source platform to organize scientific challenges. Journal of Machine Learning Research, 24 (2023) 1-6. Retrieved from https://hal.inria.fr/hal-03629462v1.
[14] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample. LLaMA: Open and Efficient Foundation Language Models, 2023. https://arxiv.org/abs/2302.13971.
[15] PaddleOCR. OCR Library. 2024. Recovered from https://pypi.org/project/paddleocr/.
[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2020. https://arxiv.org/abs/2010.11929.
[17] J. Cañete, G. Chaperon, R. Fuentes, J. Ho, H. Kang, J. Pérez. Spanish Pre-trained BERT Model and Evaluation Data, 2023. http://arxiv.org/abs/2308.02976.